Contrôle des connaissances – QCM
Sujet de l'examen

Remarques

- Durée : 2 heures.
- Une seule bonne réponse par question.
- 0,5 point par bonne réponse, 0 point par réponse indéterminée ou fausse.
- Noter les réponses sur la grille jointe.

Questions

1. À l'échelle d'un génome humain, combien retrouve-t-on de variations génétiques codantes ?

A.	Environ 1000 dont 500 non-synonymes
B.	Environ 10 000 dont 5 000 non-synonymes
C.	Environ 20 000 dont 5 000 non-synonymes
D.	Environ 20 000 dont 10 000 non-synonymes
E.	3 à 4 millions

2. Quelle affirmation est fausse ? Le séquençage nouvelle génération :

A.	Permet un débit de séquençage bien supérieur à celui du séquençage Sanger
B.	Permet un coût de séquençage par base bien inférieur à celui du séquençage Sanger
C.	A émergé en 2007-2008 avec la technologie du 454
D.	Est basé sur une construction de librairies *in vitro*
E.	Offre un taux d'erreurs par base inférieur à celui du séquençage Sanger

3. Quelle affirmation est fausse ? La technologie de séquençage d'Ion Torrent :

A.	Nécessite une étape d'amplification de PCR par émulsion
B.	Nécessite une étape d'amplification par cluster ou bridge PCR
C.	Est basée sur la libération d'un proton H+ suite à l'insertion d'un nucléotide
D.	Ne nécessite pas de nucléotides modifiés
E.	Peut être réalisée à l'aide d'un séquenceur PGM ou Proton

4. Laquelle de ces applications ne peut pas être effectuée par du DNA-Seq ?

A.	Le séquençage et la quantification de transcrits
B.	Le séquençage du génome
C.	Le séquençage de régions d'ADN ciblées
D.	La détection de variations de nombre de copies (CNVs)
E.	La détection de variations génétiques

5. Quelle étape ne fait pas partie du protocole de capture par hybridation en solution ?

A.	La fragmentation
B.	La ligation des adaptateurs
C.	L'amplification par PCR
D.	La capture des fragments d'ADN sur un support solide
E.	Le lavage pour éliminer les régions non capturées

6. La profondeur de séquençage est :

A.	Le nombre de lectures couvrant une région donnée
B.	La même chose que la couverture
C.	Obligatoirement supérieure à zéro
D.	Obligatoirement une moyenne
E.	Le nombre de bases couvertes par au moins une lecture

7. A quel format de fichier correspondent les lignes ci-dessous ?

```
@HWI-ST584:211:D15L6ACXX:4:1101:1321:2000 1:N:0:TGACCA
NATCAGAAATATGCCACACTTACTTTGAACATTCTTCCAAGGCTTGGCAGAGTGTCTGTTGAAATG
+
#41=BDDDDAD<2,AE?EEF<FDCEE>FD<EEE@EEEIIIIIIIII@D??D*099BD/>CAACEA
```

A.	Format FASTQ
B.	Format FASTA
C.	Format SAM
D.	Format BAM
E.	Format VCF

8. Quelle affirmation est fausse ? Le réalignement autour des insertions/délétions :

A.	Permet la détection de single nucleotide variants (SNVs) de façon plus spécifique
B.	Peut être réalisé avec le Genome Analysis Toolkit (GATK)
C.	Consiste à faire un réalignement sélectif de régions contenant des insertions/délétions potentielles
D.	Peut être utilisé avec un fichier VCF contenant des insertions/délétions connues
E.	Peut être réalisé avec le logiciel FastQC

9. Typiquement, la taille du fichier BAM pour un exome humain à 60X est de l'ordre de :

A.	1-10 kilooctets (ko)
B.	10-100 mégaoctets (Mo)
C.	5-10 gigaoctets (Go)
D.	1 téraoctet (To)
E.	5-10 téraoctets (To)

10. Parmi ces données produites par la plate-forme de séquençage Illumina, quel format de fichier est-il important de sauvegarder (backup) à long terme ?

A.	Les fichiers d'images
B.	Les fichiers d'intensités
C.	Les fichiers FASTQ
D.	Les fichiers SAM
E.	Les fichiers BED

11. Le logiciel FileZilla :

A.	Permet l'échange de fichiers entre un client et un serveur à distance
B.	Est un interpréteur de commande distant
C.	Remplit les mêmes fonctions que Putty
D.	Est un logiciel pour un environnement graphique distant
E.	Ne fonctionne que sous Linux

12. Sous Linux, laquelle de ces commandes doit être utilisée pour se rendre dans le répertoire /work/c-shd/shared/ ?

A.	pwd /work/c-shd/shared/
B.	mv /work/c-shd/shared/
C.	rm /work/c-shd/shared/
D.	cd ~
E.	cd /work/c-shd/shared/

13. Sous Linux, la commande pour afficher le mode d'emploi de la commande « cd » est :

A.	man cd
B.	pwd cd
C.	ls cd
D.	ls -l cd
E.	mkdir cd

14. Sous Linux dans l'exemple ci-après, le fichier « myfile.txt » peut être :

```
-rw-rw---- 1 je8762ri gad myfile.txt
```

A.	Lu et modifié par tous les utilisateurs
B.	Lu et modifié par son propriétaire et les membres du groupe du propriétaire
C.	Lu et modifié par son propriétaire seulement
D.	Lu par tous les utilisateurs
E.	Exécuté par tous les utilisateurs

15. Le fichier « hg19.fa » est le génome de référence humain en format FASTA. Sous Linux, la commande ci-dessous permet de :

```
grep '>' hg19.fa
```

A.	Faire apparaître toutes les lignes qui ne commencent pas par le symbole >
B.	Compter le nombre de nucléotides contenus dans ce fichier
C.	Compter le nombre de chromosomes de ce fichier
D.	Afficher la liste des chromosomes de ce fichier
E.	Convertir le fichier hg19.fa en format FASTQ

16. Sous Linux, la commande suivante permet de :

```
awk -F "\t" '{print $1":"$2"-"$3}' myfile.bed > myfile.list
```

A.	Transformer le fichier myfile.bed (format BED) en fichier LIST (myfile.list)
B.	Transformer n'importe quel ficher BED dans le répertoire de travail en fichier LIST
C.	Transformer le fichier myfile.list (format LIST) en fichier BED (myfile.bed)
D.	Transformer n'importe quel fichier LIST dans le répertoire de travail en fichier BED
E.	Renommer le nom du fichier myfile.bed par myfile.list

17. Sous Python, quel est le type de l'élément ci-dessous :

```
[1, 2, 3, 4, 5]
```

A.	Integer
B.	String
C.	List
D.	Tuple
E.	Boolean

18. Sous Python, suite aux 3 commandes ci-dessous, « a » est :

```
a = 'actgactg'
a = len(a)
a = 8
```

A.	Une variable dont la valeur assignée est `'actgactg'`
B.	Une variable dont la valeur assignée est de type String
C.	Une variable dont la valeur assignée est une liste : `['actgactg', len(a), 8]`
D.	Une variable dont la valeur assignée est 8
E.	Aucun des choix précédents, a n'est pas une variable

19. Sous python, la boucle ci-dessous permet de fournir :

```
for x in range(0, 101):
      print x
```

A.	Tous les nombres entiers de 0 à 100
B.	Tous les les nombres entiers pairs de 0 à 100
C.	Tous les nombres entiers impairs de 0 à 100
D.	100 fois la lettre x
E.	Des nombres croissants à l'infini

20. Sous python, le script ci-dessous permet de :

```
import sys
myfile = sys.argv[1]

for line in open(myfile):
      if line.startswith('#'):
            print line.strip()
      else:
            coord = line.strip().split()
            chrom = coord[0]
            prox = int(coord[1])
            dist = int(coord[2])
            print "%s:%s-%s" % (chrom, prox + 1, dist)
```

A.	D'imprimer toutes les lignes d'un fichier BED, sauf celles qui commencent par le signe #
B.	De n'imprimer que les lignes d'un fichier BED qui commencent par le signe #
C.	De produire un nouveau fichier BED en ajoutant +1 à la position proximale
D.	De convertir n'importe quel fichier BED en format LIST (chrom:prox-dist) et d'imprimer le header
E.	De convertir n'importe quel fichier BED en format LIST (chrom:prox-dist) sans imprimer le header

21. Quelle affirmation est fausse ? Le logiciel FastQC permet :

A.	D'évaluer la qualité des données de séquençage brutes
B.	D'effectuer un alignement des séquences à un génome de référence
C.	D'évaluer le score de qualité des bases en fonction du cycle
D.	D'estimer la représentation des nucléotides A, C, G et T en fonction du cycle
E.	D'estimer le pourcentage de GC en fonction du cycle

22. Quelle affirmation est fausse ? Le format de fichier SAM :

A.	Est le format générique de variations génétiques
B.	Est le format générique d'alignement de séquences
C.	Est un fichier texte tabulé
D.	Convient aux lectures courtes et longues
E.	Est un sigle qui veut dire « Sequence Alignment/MAP »

23. Quelle étape de manipulation de données de DNA-Seq ne s'applique pas à l'enrichissement par amplicons ?

A.	L'alignement des séquences à un génome de référence
B.	La suppression des duplicats de PCR
C.	Le réalignement autour des insertions/délétions
D.	La recalibration de la qualité des bases
E.	La conversion du format SAM en format BAM

24. La commande suivante permet :

```
samtools index sample.bam
```

A.	De créer un fichier index du fichier « sample.bam » à l'aide de Samtools
B.	De créer un fichier index du fichier « sample.bam » à l'aide de Picard
C.	D'indexer les duplicats de PCR du fichier « sample.bam » à l'aide de Samtools
D.	De compter le nombre de duplicats de PCR du fichier « sample.bam » à l'aide de Samtools
E.	De visualiser le header du fichier « sample.bam » à l'aide de Samtools

25. La commande suivante du GATK permet :

```
GATK -T UnifiedGenotyper -R hg19.fa -I sample.bam -o sample.vcf
```

A.	D'identifier des variations génétiques dans l'échantillon « sample.bam »
B.	De filtrer les variations de mauvaise qualité du fichier « sample.vcf »
C.	De créer un ficher index pour l'échantillon « sample.bam »
D.	De visualiser en parallèle le fichier BAM « sample.bam » et ses variations dans « sample.vcf »
E.	De créer un fichier BAM à partir d'un fichier VCF

26. Quelle option du UnifiedGenotyper ne figure pas dans le bloc d'options ci-dessous :

```
-mbq 20 \
-minIndelCnt 2 \
-minIndelFrac 0.05 \
-stand_call_conf 30.0 \
```

A.	Le seuil de qualité des bases
B.	Le pourcentage minimal d'insertions/délétions
C.	Le score de confiance minimal pour identifier une variation
D.	Le nombre minimal de lectures avec une insertion/délétion donnée
E.	Un fichier de dbSNP pour renseigner la colonne ID du fichier VCF

27. Quelle approche bioinformatique ne permet pas d'identifier des CNVs à partir de données de DNA-Seq ?

A.	L'assemblage de novo
B.	La méthode de « split-read »
C.	Le réalignement autour des insertions/délétions
D.	La différence de profondeur relative
E.	L'identification d'anomalies d'appariement des lectures

28. Lequel de ces outils ne permet pas l'annotation de variations génétiques ?

A.	SeattleSeq Annotation
B.	Annovar
C.	SnpEff
D.	Sequence Variant Analyzer (SVA)
E.	Le Burrows-Wheeler Aligner (BWA)

29. D'après l'extrait d'un fichier VCF ci-dessous :

```
FORMAT            Patient
GT:AD:DP:GQ:PL    0/1:10,8:18:99:204,0,244
```

A.	Le patient est hétérozygote
B.	Le patient est homozygote pour l'allèle alternatif
C.	Le patient est homozygote pour l'allèle de référence
D.	La profondeur de séquençage est de 99
E.	10 lectures supportent l'allèle mutant

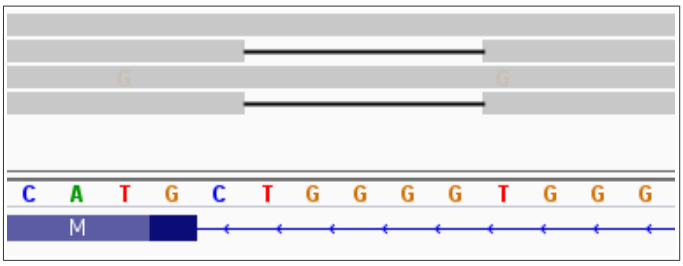
30. Vous analysez les variations génétiques d'exomes d'un trio (proband, père, mère). Quelle approche est susceptible de produire la liste la plus courte de variations candidates ?

A.	Ne garder que les variations avec un score de qualité ≥ 30
B.	Ne garder que les variations non-synonymes
C.	Ne garder que les variations *de novo*
D.	Ne garder que les variations absentes de l'Exome Variant Serveur (6,500 exomes individuels)
E.	Ne garder que les variations présentes à une fréquence < 0,1 % dans dbSNP138

31. Le logiciel Integrative Genomics Viewer (IGV) ne permet pas :

A.	De visualiser des lectures alignées en format BAM
B.	D'indexer un fichier VCF
C.	De visualiser un fichier VCF
D.	De visualiser un fichier BED
E.	D'identifier des variations génétiques à l'échelle d'un génome

32. La variation génétique ci-dessous visualisée à l'aide d'IGV est :



A.	Une délétion de 4 pb
B.	Une délétion de 5 pb
C.	Une insertion de 4 pb
D.	Une insertion de 5 pb
E.	Un changement d'un seul nucléotide (SNV)

33. COSMIC :

A.	Est une base de données de variations somatiques dans des cancers humains
B.	Est une base de données de variations germinales
C.	Ne contient que des variations génétiques réelles et bien annotées
D.	Est un logiciel d'annotation de variations génétiques
E.	Est une base de données de génomes extraterrestres

34. Quelle affirmation est fausse ? La base de données dbSNP :

A.	Ne contient que des variations génétiques non-pathogènes depuis la version 131
B.	Contient des SNVs et des insertions/délétions
C.	Contient des millions de variations génétiques
D.	Est une base de données généraliste de variations génétiques
E.	Est libre d'accès

35. Vous travaillez en recherche sur une pathologie génétique rare et sporadique. De manière générale, si vous identifiez une variation *de novo* d'un gène qui n'a jamais associé à cette pathologie chez un patient, quelle est la première chose que vous faites ?

A.	Vous écrivez un rapport au médecin traitant pour l'informer de votre découverte
B.	Vous effectuer des analyses fonctionnelles afin de confirmer le caractère pathogène de la variation
C.	Vous séquencez le gène chez d'autres patients avec le même phénotype pour confirmation
D.	Vous n'en faites rien, on trouve des variations *de novo* chez tout le monde
E.	Vous testez les grands-parents pour vous assurer que la variation n'a pas sauté une génération

36. Quelle affirmation est fausse ? En recherche en génétique :

A.	Le séquençage d'exomes par trios combiné à des études de liaison facilite l'identification de variations *de novo*
B.	Le séquençage du génome permet d'identifier un éventail de variations plus large que l'exome
C.	Le séquençage du génome permet d'identifier à peu près autant de variations codantes que l'exome
D.	Le mode de transmission prédit d'une pathologie génétique donnée a un impact sur la stratégie d'identification du gène responsable
E.	Il est prudent de visualiser les variations candidates sur IGV avant confirmation par Sanger

37. Les algorithmes de prédiction *in silico* de SNVs sont :

A.	À éviter systématiquement car ils ne sont pas fiables
B.	Toujours basés sur la conservation inter-espèces
C.	Toujours limités à l'analyse de variations non-synonymes
D.	Souvent plus performants pour les pertes de fonction que pour les gains de fonction
E.	Efficaces seulement si la variation génétique a déjà été rapportée

38. En génétique médicale, une variation est considérée comme pathogène si :

A.	Elle est considérée comme pathogène par des prédictions *in silico*
B.	Elle est *de novo* et touche un gène associé à la pathologie étudiée
C.	Elle est tronquante et touche un gène associé à la pathologie étudiée
D.	Elle a été précédemment rapportée comme pathogène dans la pathologie étudiée
E.	Elle n'a jamais été identifiée dans les bases de données publiques de variations génétiques

39. Le Genome Browser Ensembl :

A.	Remplit les mêmes fonctions qu'IGV
B.	Ne contient que des données sur le génome humain
C.	Est une collaboration entre le BGI et Celera Genomics
D.	Ne permet pas de visualiser ses propres données
E.	A son propre système d'identifiants de gènes, de transcrits et de peptides

40. L'outil du UCSC Genome Browser permettant d'obtenir de nombreuses données d'annotation en format texte est :

A.	ExonPrimer, basé sur l'algorithme de Primer3
B.	Table Browser
C.	Custom Tracks
D.	BLAST-like Alignement Tool (BLAT)
E.	Inexistant, l'utilisation de Galaxy est nécessaire pour effectuer cette tâche